Available online at www.sciencedirect.com

ScienceDirect

www.sciencedirect.com/science/journal/16720229

GENOMICS PROTEOMICS & BIOINFORMATICS

Article

The Most Redundant Sequences in Human CpG Island Library Are Derived from Mitochondrial Genome

Ximiao He[1,2#], Shu Tao[1,2#], Jing Jin[3], Songnian Hu[1*], and Jun Yu[1*]

[1]*CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China;*
[2]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China;*
[3]*Department of Biology, University of Science and Technology of China, Hefei 230027, China.*

Genomics Proteomics Bioinformatics 2010 Jun; 8(2): 81-91. DOI: 10.1016/S1672-0229(10)60009-5

Abstract

An altered pattern of epigenetic modifications, such as DNA methylation and histone modification, is critical to many common human diseases, including cancer. Recently, mitochondrial DNA (mtDNA) was reported to be associated with tumorigenesis through epigenetic regulation of methylation patterns. One of the promising approaches to study DNA methylation and CpG islands (CGIs) is sequencing and analysis of clones derived from the physical library generated by methyl-CpG-binding domain proteins and restriction enzyme MseI. In this study, we observed that the most redundant sequences of 349 clones in a human CGI library were all generated from the human mitochondrial genome. Further analysis indicated that there was a 5,845-bp DNA transfer from mtDNA to chromosome 1, and all the clones should be the products of a 510-bp MseI fragment, which contained a putative CGI of 270 bp. The 510-bp fragment was annotated as part of cytochrome c oxidase subunit II (COXII), and phylogenetic analysis of homologous sequences containing COXII showed three DNA transfer events from mtDNA to nuclear genome, one of which underwent secondary transfer events between different chromosomes. These results may further our understanding of how the mtDNA regulates DNA methylation in the nucleus.

Key words: human, DNA methylation, CpG islands, nuclear mitochondrial DNA, molecular phylogeny

Introduction

DNA methylation is a critical biochemical modification of eukaryotic DNA involved in various biological processes including gene silencing, chromosomal structure stabilization, X-chromosome inactivation, imprinting, and cell differentiation (*1-9*). In mammals, DNA methylation mainly occurs at the fifth carbon position of the cytosine in a CpG context, and this biochemical process is owing to the activity of DNA methyltransferases (DNMTs) (*10, 11*). The dinucleotide CpG is remarkably under-represented in the human genome with only about 20% of the expected frequency. However, there are many genomic regions that contain a much high frequency (about 10 times higher than the average of genome) of CpG dinucleotides, and these regions are called CpG islands (CGIs) (*12*). The usual formal defination of a CpG island is a region at least 200 bp in length, with a GC percentage greater than 50% and an observed/expected CpG ratio greater than 0.6 (*13*).

# Equal contribution.
*Corresponding authors.
E-mail: husn@big.ac.cn; junyu@big.ac.cn

Recently, a stricter rule (>500 bp, GC content >55%, and CpG ratio >0.65) of CpG island prediction was suggested in order to exclude other GC-rich genomic sequences such as Alu repeats (*14*).

The CGIs are traditionally thought to be unmethylated, even some of them were found to be hypermethylated in the imprinted genes (*15*). Considering this property of CGIs, and that unmethylated CpG sites can be methylated *in vitro*, Cross *et al* introduced an approach to purify CGIs using a methylated DNA binding (MDB) column (*16*). Heisler *et al* analyzed the CGI clones derived from a physical library generated by MDB and MseI, and suggested that the clones are representatives of CGIs annotated on the human genome, and there may be value in continuing to isolate clones from the library (*17*). The CGIs typically occur at or near the transcription start site of genes, particularly housekeeping genes, in vertebrates. The methylation of CGIs is involved in the regulation of gene expression, which plays an important role in disease development including tumorigenesis (*18, 19*). To better understand the interplay of CGI methylation and cancer, He *et al* continued to isolate and sequence the clones derived from the same CGI library, and the sequences were deposited in the database of MethyCancer (*20*). Based on the comprehensive analysis of these CGI clones, we observed that the most redundant sequences of 349 clones were generated from the human mitochondrial genome (mtDNA).

The nuclear mitochondrial DNA, first denoted as "NUMT" in cat by Lopez *et al* (*21*), refers to DNA segment that has been transferred from mitochondrial genome to nuclear genome. This phenomenon has been observed in diverse eukaryotes such as human, mouse, rat, rice, *Arabidopsis* and insects, with a large NUMT number and size variation across species (*22*). Recently, mtDNA was found to be associated with tumorigenesis through epigenetic regulation of methylation patterns. Xie *et al* (*23*) studied the effect of mtDNA depletion on cancer progression and found that mtDNA depletion promotes cancer progression through activating hypermethylation pattern of cancer-associated genes' promoter CpG islands, and this activation was achieved through the induction of DNMT1. Similarly, Smiraglia *et al* (*24*) investigated whether mtDNA copy number variation, a feature of many human tumors, can affect methylation changes in the nucleus, and found that methylation pattern is reversible for a number of genes following depletion and restoration of mtDNA.

In this study, we reported a DNA transfer of about 6k mtDNA (namely $NUMT^{ND\text{-}COX}$) to chromosome 1, which covered all of the most redundant 349 clones. Further analysis of $NUMT^{ND\text{-}COX}$ indicates that all the clones should be the products of an MseI fragment with the length of 510 bp ($NUMT^{ND\text{-}COX}$: 3,841-4,110), which contained a putative CGI of 270 bp. Phylogenetic analysis of homologous sequences containing cytochrome c oxidase subunit II (COXII) allows us to detect three DNA transfer events, and one of these events underwent possible secondary interchromosomal transfer events. This observation may provide us a clue for further understanding of the roles of mtDNA in regulation of DNA methylation in the nucleus.

Results

Genomic mapping of CGI library clones

Heisler *et al* (*17*) analyzed and compared the CGI clones of 12k set (12,192 clones) deposited at the Wellcome Trust Sanger Institute, and the 9k set (8,554 clones) isolated from the same CGI library (*25*) by the Huang laboratory, and found that there was only a small degree of overlap between the two sets, with only 753 common genomic loci of the total 9,595. While in 17,606 sequences obtained from MethyCancer, 17,068 sequences were aligned to the human genome using BLAT/BLAST (*26*). Clones sharing the overlapped genome locations were clustered into 10,648 distinct genomic loci, with the redundancy of 37.61%.

We combined the three datasets of BIG18K (the 17,606 clones sequenced by Beijing Institute of Genomics, CAS), Sanger12K (the 12,192 clones deposited at Sanger) and UHN8K (the 8,373 clones downloaded from UHN). Overall, the 35,602 clones mapped to the genome were clustered into 18,240 genomic loci. The number of distinct genomic loci of BIG18K was 7,932 (74.49%), and the number of common loci of the three sets was only 913 (**Figure 1A**), which implies that the majority of loci of clones

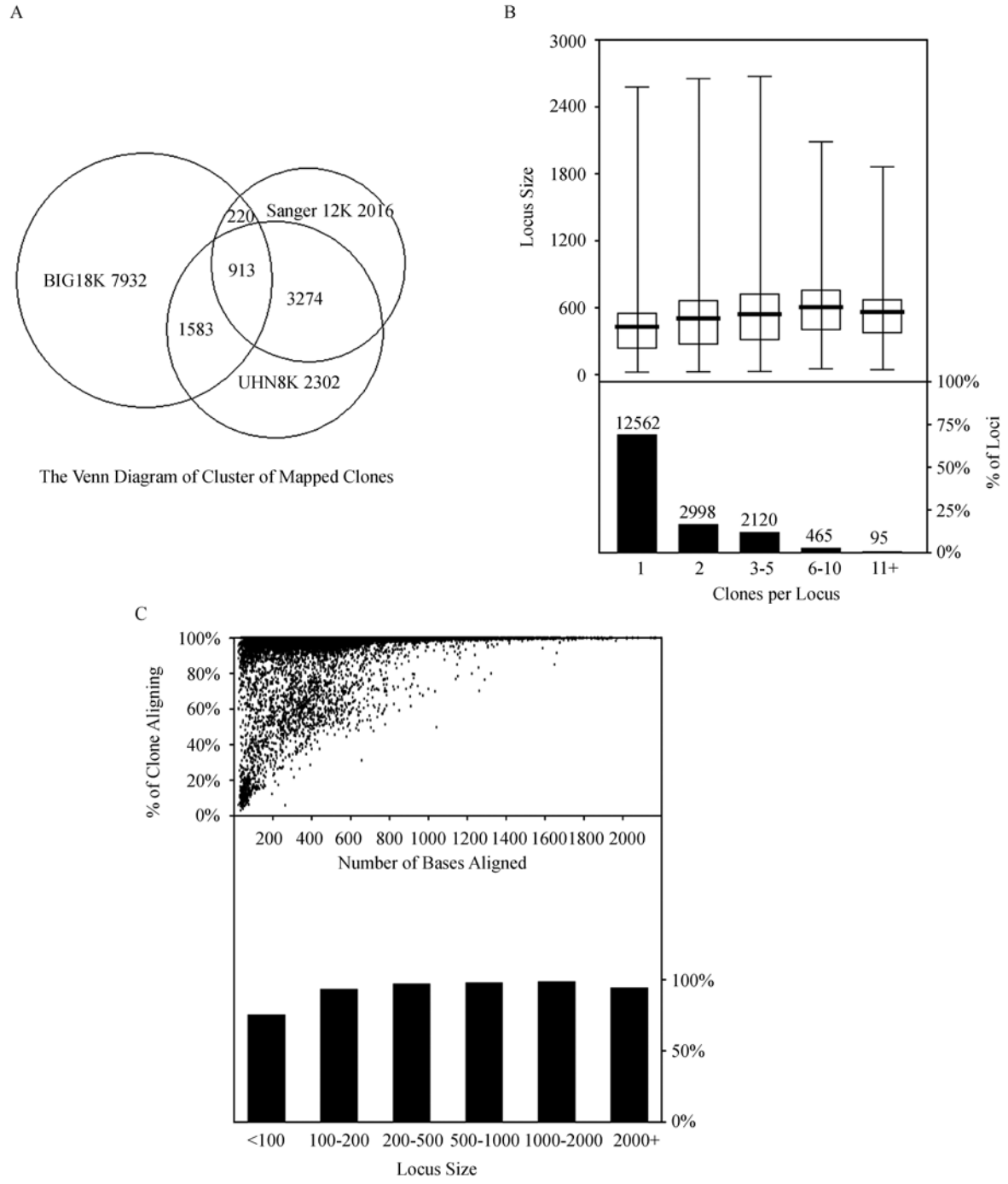

Figure 1 Genomic loci defined by the aligned clones of combined dataset of CGI clones from BIG18K, Sanger12K and UHN8K. **A.** The Venn diagram of loci defined by sequenced CGI library clones of the three datasets. Only 913 loci are common. 7,932, 2,016 and 2,302 loci are distinct for BIG, Sanger and UHN datasets, respectively. **B.** Clone component of genomic loci. Each genomic locus is defined by one or more alignments of CGI clones sequenced. In the upper panel, the loci size ranged from 23 to 2,673 bp, with a mean length of 460.76 bp. In the lower, the majority of the loci (12562/18240) are defined by single non-redundant clones. The others are defined by redundant ones, with 2,998 loci by 2 clones, 2,120 by 3-5, 465 by 6-10, and 95 by more than 11 clones. **C.** Percentage of sequence aligning. In the upper panel, the percentage of alignment of clone length is plotted against the total number of bases aligned. The shorter alignments mostly are generated from partial alignments. In the lower panel, the percentage of sequence alignment for loci of various size ranges is shown. The smaller loci (<100 bp) are generated mostly from partial alignments, while loci of 200 bp and greater are generally resulted from nearly complete alignments.

BIG18K are distinct, and the clones sequenced by BIG are complementary to Sanger12K and UHN8K clone sets. The total redundancy of the combined set was 48.77%. There were 12,562 loci (68.87%) defined by a single clone (**Figure 1B**), which accounted for only 42.87% of all clones aligned. About 38.01% of clones had a low degree of redundancy (2-5 per locus), while 7.19% were highly redundant (11+ per locus). The genomic loci ranged from 23 to 2,673 bp, with a mean length of 460.76 bp, while the lengths of clones mapped to the genome were between 22 and 2,998 bp, with the mean of 503.57 bp. The loci of 200 bp and greater are generally resulted from nearly complete alignments, while the smaller loci (<100 bp) are generated mostly from partial alignments (**Figure 1C**).

The most redundant sequences are from mitochondria

Among 95 loci with the high redundancy (with more than 11 clones per locus), we are very interested in the locus HsCGICLT000005 with the highest redundancy, where 349 clones are fairly well mapped to a region of 640 bp (557,995-558,634) in chromosome 1 (**Table 1**). We selected UHNhscpg0011553 as the representative clone, which had covered the most part of the locus (512/640). Blast of UHNhscpg0011553 showed that this clone had best hit on human mitochondria and chromosome 1. The max score and identity of the hit to mtDNA (NC_001807.4) was 921 and 511/512, respectively, while hit to chromosome 1 was 884 and 503/512 (**Table S1**), which implied that the clone UHNhscpg0011553 should be derived from mitochondria. To identify the sources of all the 349 clones, that is, mitochondria or nuclear genomes, we blast them and picked up the best hit to mtDNA and chromosomes. To our surprise, all of the 349 clones had the better hit to mitochondria than chromosome 1, according to the max score and identity of blast (**Figure 2**), which implied that all the clones should be sourced from mtDNA, while none of them was sampled from the chromosome 1.

Table 1 The top 20 genomic loci with high redundancy

Genomic loci	Chromosome	Start position	End position	Length	No. of clones
HsCGICLT000005	Chr01	557,995	558,634	640	349
HsCGICLT014311	Chr16	33,871,368	33,871,672	305	158
HsCGICLT005628	Chr05	134,286,927	134,287,597	671	91
HsCGICLT016746	Chr20	13,095,958	13,096,001	44	86
HsCGICLT000868	Chr01	121,185,813	121,186,957	1,145	82
HsCGICLT017988	Chr23	108,184,004	108,184,491	488	74
HsCGICLT002468	Chr02	132,729,663	132,730,140	478	60
HsCGICLT014330	Chr16	44,960,953	44,961,298	346	59
HsCGICLT005630	Chr05	134,288,999	134,289,975	977	55
HsCGICLT000006	Chr01	559,565	560,167	603	47
HsCGICLT002470	Chr02	132,741,922	132,742,484	563	47
HsCGICLT016153	Chr19	32,423,791	32,424,421	631	45
HsCGICLT006217	Chr06	26,888,125	26,888,564	440	42
HsCGICLT007524	Chr07	61,607,698	61,608,044	347	42
HsCGICLT006579	Chr06	58,256,526	58,256,963	438	40
HsCGICLT011044	Chr11	84,872,658	84,872,952	295	40
HsCGICLT009855	Chr10	41,720,317	41,720,723	407	39
HsCGICLT005631	Chr05	134,290,504	134,290,919	416	38
HsCGICLT007523	Chr07	61,606,003	61,607,317	1,315	37
HsCGICLT014308	Chr16	33,870,097	33,870,571	475	37
HsCGICLT014306	Chr16	33,863,874	33,864,181	308	35

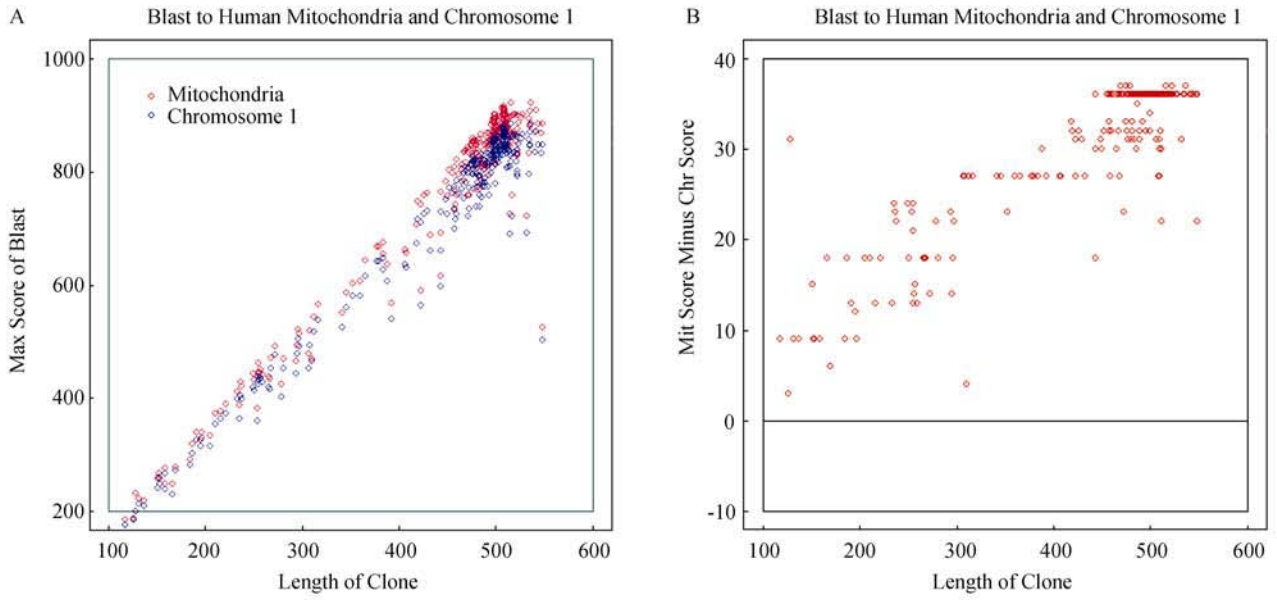

Figure 2 Blast result of 349 clones to human mitochondria and chromosome 1. **A.** Max score of the blast is plotted against the length of each clone. Red circle indicates the blast result to human mitochondria, and blue circle shows blast to chromosome 1. **B.** The blast max score of mitochondria minus max score of chromosome 1 is plotted against the clone length. All the circles are plotted above the line of score 0, which means all the clones have the better hits on the mitochondria than chromosome 1.

Among the 349 clones, Clone002352 is the longest one, with the length of 915 bp; blast result showed that its alignment to mtDNA was partial (1-474), and the rest (475-915) was mapped to chromosome 9. Like Clone002352, Clone019324 in a length of 833 bp was partially aligned to mtDNA (1-491), and the rest (492-833) was mapped to chromosome X. Unlike the above two clones, Clone008551_Connected had an insertion of 54 bp (304-357) when compared with the mtDNA and other clones (**Figure S1**).

GC content and CpG ratio of the 349 clones

To evaluate the potential of the clones to be CpG islands, we calculated the GC content and CpG dinucleotide frequency. CpG ratio of the majority of clones, from 0.8 to 1.0, was much higher than the CGI criteria of CpG ratio of 0.6, while GC content (46%-48%) was lower than the criteria of 50% (**Figure 3**). The average of CpG ratio was 0.92 and the median was 0.91, the minimal was 0.53 (Clone017492, 159 bp) and the maximal was 1.33 (Clone026096_5, 185 bp), while the average of GC content was 46.82% and the median was 46.86%, the minimal was 38.83% (UHNhscpg0002733, 515 bp) and the maximal was 51.87% (Clone021408_5, 187 bp). As shown in Figure 3, dispersed spots are the shorter clones (<475 bp, Figure 3A) and longer ones (>512 bp, Figure 3F), respectively.

CGI prediction and MseI fragments of NUMT$^{ND\text{-}COX}$

To identify the homologous region of mtDNA that includes the locus HsCGICLT000005 on chromosome 1, we extended the sequence of locus by 20 kb (10 kb of upstream and downstream, respectively), and searched the homology all over the whole mtDNA by blast of bl2seq. A mitochondrial homologous region of 5,845 bp (3,911-9,755) was determined, which covered the genes of ND1, ND2, COX1, COX2, and COX3, labeled as NUMT$^{ND\text{-}COX}$. Then we carried out the CGI prediction of NUMT$^{ND\text{-}COX}$ using CpGi130 (*27*) and 6 putative CGIs were predicted (**Table 2**). Among them, NUMT$^{ND\text{-}COX}$CGI5 with the length of 270 bp (NUMT$^{ND\text{-}COX}$: 3,841-4,110) had the greatest CpG ratio of 1.05.

Considering that the putative CGIs may be digested by restriction enzyme MseI when the library of CGI clones was constructed, we searched all of the MseI

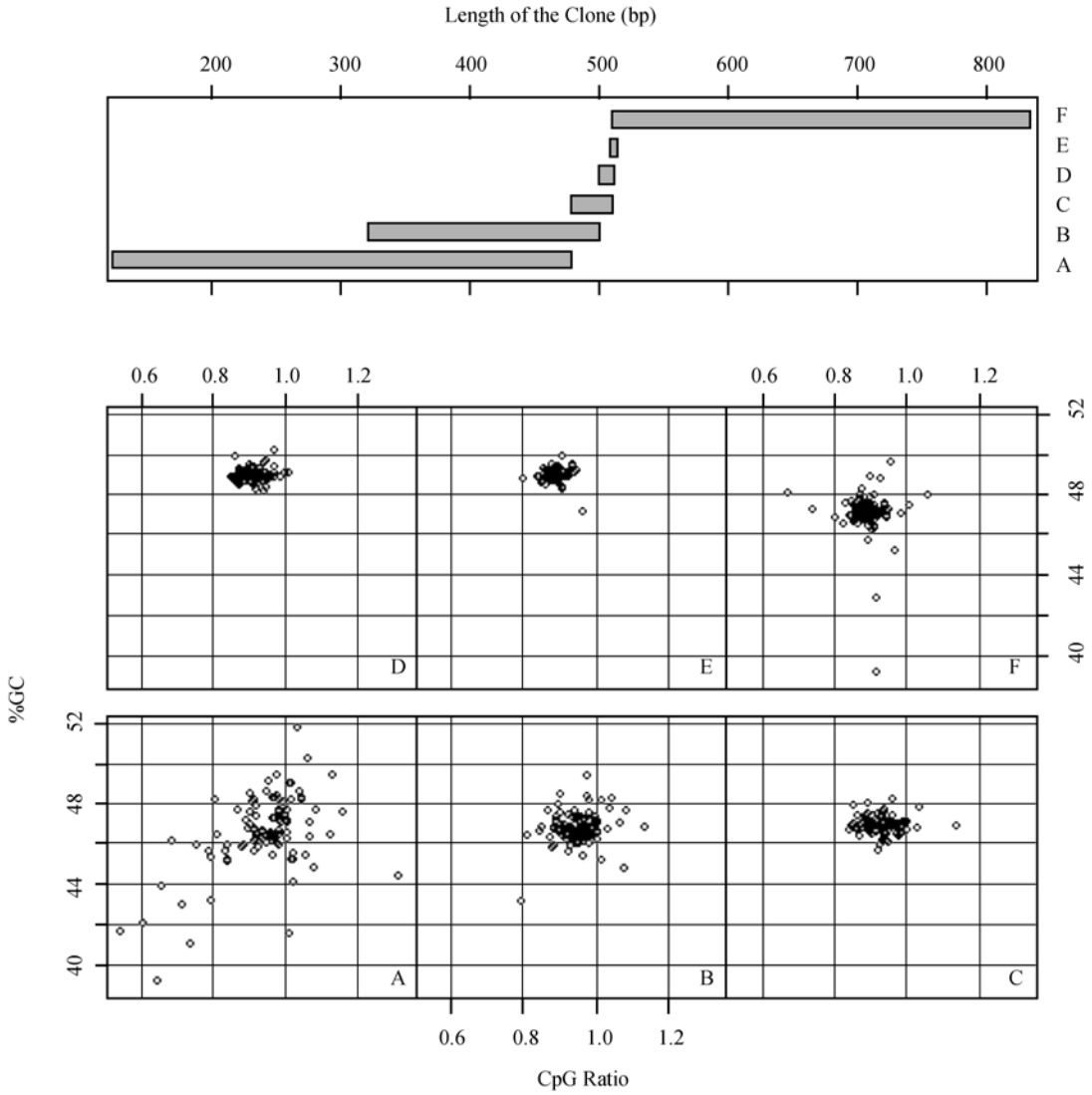

Figure 3 GC content and CpG ratio of 349 clones. With the step of 50 clones and window of 100 clones, the GC content and CpG ratio of 349 clones are plotted in six panels: **A**. The first 100 clones ordered by length from 117 to 475 bp. **B**. 100 clones (51-150) with length from 317 to 496 bp. **C**. 100 clones (101-200) with length from 476 to 506 bp. **D**. 100 clones (151-250) with length from 497 to 508 bp. **E**. 100 clones (201-300) with length from 506 to 511 bp. **F**. The last 99 (251-349) clones with length from 512 to 915 bp. The CpG ratio is calculated by (#CpG×#NT)/(#C×#G), where #CpG is the number of CpG dinucleotide in the clone, #NT is the length of clone, #C and #G is the number of C and G in the clone, respectively.

Table 2 CpG island prediction of NUMT$^{ND\text{-}COX}$ using CpGi130

CpG island No.	Start position	End position	GC%	CpG ratio	Length
NUMT$^{ND\text{-}COX}$CGI1	443	642	50	0.793	200
NUMT$^{ND\text{-}COX}$CGI2	1,821	2,186	50	0.774	366
NUMT$^{ND\text{-}COX}$CGI3	2,200	2,399	50	0.6	200
NUMT$^{ND\text{-}COX}$CGI4	2,508	2,713	50	0.76	206
NUMT$^{ND\text{-}COX}$CGI5	3,841	4,110	50	1.05	270
NUMT$^{ND\text{-}COX}$CGI6	5,349	5,548	51.5	0.627	200

recognition sequences (TTAA) on the whole $NUMT^{ND\text{-}COX}$. As a result we obtained 29 MseI fragments, among which 8 MseI fragments had a length of over 200 bp (**Table 3**). Comparing with the CGI predictions, we observed that three of them ($NUMT^{ND\text{-}COX}$CGI2, $NUMT^{ND\text{-}COX}$CGI3, and $NUMT^{ND\text{-}COX}$CGI4) located in the largest fragment (1,610 bp) of $NUMT^{ND\text{-}COX}Frag^{MseI}16$, $NUMT^{ND\text{-}COX}$CGI5 lied in $NUMT^{ND\text{-}COX}Frag^{MseI}19$, and $NUMT^{ND\text{-}COX}$CGI6 was in the 3'-end fragment of $NUMT^{ND\text{-}COX}$, while the prediction of $NUMT^{ND\text{-}COX}$CGI1 was interrupted by MseI recognition sequence. Homologous searches of the 349 clones indicated that all the clones should be the products of $NUMT^{ND\text{-}COX}Frag^{MseI}19$ with the length of 510 bp ($NUMT^{ND\text{-}COX}$: 3,841-4,110), which contained a putative CGI of $NUMT^{ND\text{-}COX}$CGI5.

Phylogenetic analysis of the $NUMT^{ND\text{-}COX}$

As mentioned above, all the 349 clones were mapped exclusively to a single region (around 510 bp) in mitochondrial genome and chromosome 1, respectively, and this 510-bp region was annotated as part of COXII using Blastx. Therefore, we collected 6 NUMTs (Chr1 6kb and Chr2, 4, 7, 11, 17) containing COXII homologous sequences and 20 mitochondrial sequences containing COXII from a wide range of species and subjected them to phylogenetic analysis (**Figure 4**). We were able to identify three insertion events that occurred at different times during evolution. The most ancient insertion event happened during the split of New World monkeys (spider monkey) and Old World monkeys (macaque, grivet, baboon, *etc.*), and this insertion sequence underwent several duplication events to generate 4 NUMTs (Chr2, 4, 7 and 11) with the evolution of genome, which was consistent with previous research by Tourmen *et al* (*28*). The other two primary insertion events were more recent occurred on chromosome 1 (Chr1 6kb) and chromosome 17 (Chr17).

Discussion

BLAT was our first choice to align clones to the genome because of its speed and sensitivity. The insertion and deletion were considered in the result, which assumed that there were introns in the genomic sequence. However, the CGI library clones were generated directly from the genome by the methylation-sensitive restriction enzyme MseI, so there should be no introns in the CGI clones. For this reason, if the length of insertion/deletion was more than 50 bp or over 10% of the clone, we picked it out to do BLAST. Based on the position information of alignment, all the clones mapped to the genome are clustered to define the genomic loci. Using the genomic loci, we identified the redundancy of the CGI library clones. There were a few clones mapped to multiple genomic loci because of duplication of genome fragment, and we only chose the best one. There should be enough redundant clones to define each duplicated genomic locus, even each clone was located only once on the genome arbitrarily. With more clones sequenced and aligned to the genome, the information of duplicated loci would be obtained.

Table 3 MseI fragment of $NUMT^{ND\text{-}COX}$ (Recognition Site: T|TAA, >200 bp)

MseI fragment	Start position	End position	Length	Note
$NUMT^{ND\text{-}COX}Frag^{MseI}16$	1,783	3,393	1,610	Cover $NUMT^{ND\text{-}COX}$CGI2, 3, 4
$NUMT^{ND\text{-}COX}Frag^{MseI}19$	3,674	4,184	510	Cover $NUMT^{ND\text{-}COX}$CGI5
$NUMT^{ND\text{-}COX}Frag^{MseI}1$	0	404	404	5′-end, length >404 bp
$NUMT^{ND\text{-}COX}Frag^{MseI}26$	4,541	4,840	299	
$NUMT^{ND\text{-}COX}Frag^{MseI}3$	562	846	284	
$NUMT^{ND\text{-}COX}Frag^{MseI}17$	3,393	3,646	253	
$NUMT^{ND\text{-}COX}Frag^{MseI}30$	5,239	5,485	246	3′-end, >246 bp, Cover $NUMT^{ND\text{-}COX}$CGI6
$NUMT^{ND\text{-}COX}Frag^{MseI}4$	846	1,068	222	

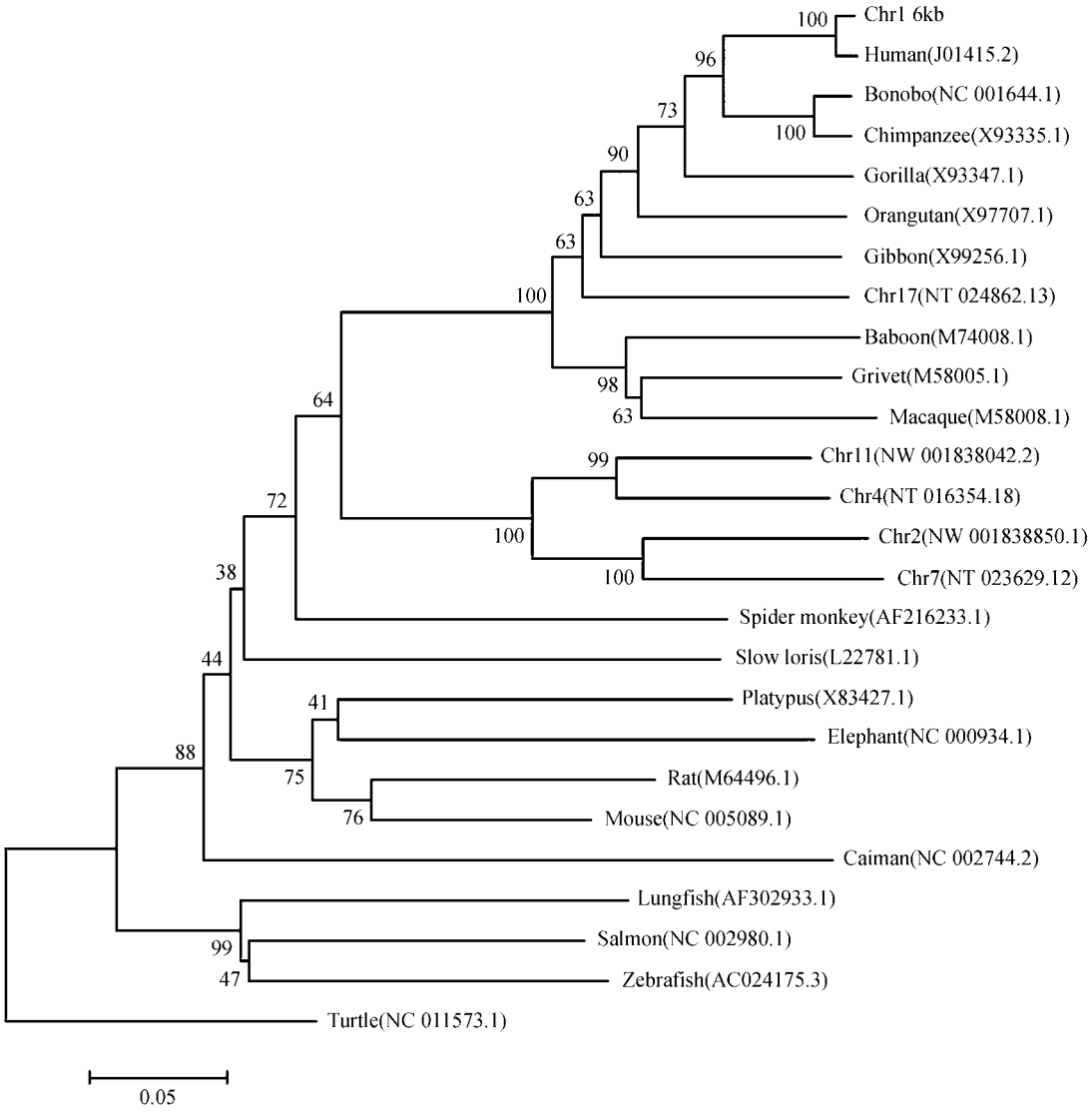

Figure 4 Phylogenetic analysis of homologous nuclear and mitochondrial COXII sequences. Two primary DNA transfers (Chr1 6kb and Chr17 NUMTs) from mitochondria to nuclear genome are shown. Other four NUMTs (Chr11, 4, 2 and 7) in a single cluster reflect secondary interchromosomal transfer events.

It is worth noting that the CGI locus is not equal to CGI itself but the tag sequence to CGI. Firstly, the library inserts are genomic MseI fragments, and the enzyme that recognizes the sequence of TTAA tends to cut outside islands, but without promise of just outside. So the CGI library clones do not necessarily have the same properties as CpG islands (*16*). Secondly, for some CGI library clones, even they are the CGI themselves, the end parts of the clones may be cutoff because of sequencing quality, and the sequence may be partially aligned to the genome. For example, some CGI loci are less than 100 bp, and the GC contents are no more than 50%. However, the major property of position to the gene transcript start site changes little (unpublished data, not shown), so we can still investigate the loci position to identify which gene the CGI regulates.

Interestingly, according to BLAT or BLAST result, there were 349 clones clustered into chromosome 1, but they have better hit score and identity to mtDNA than chromosome 1, which implied that all of them were derived from mitochondrial genome. Using blast of bl2seq, we determined a DNA transfer of 5,845 bp from mtDNA ($NUMT^{ND\text{-}COX}$) to chromosome 1. CGI prediction and MseI fragment analysis indicated that the 349 clones should be sourced from a 510-bp MseI

fragment of $NUMT^{ND\text{-}COX}Frag^{MseI}19$, which had a 270-bp putative CGI of $NUMT^{ND\text{-}COX}$CGI5. The putative CGI (with the CpG ratio of 1.05) may account for the high affinity to methyl-CpG-binding domain proteins (MBDs) and the structure of DNA may be suitable for PCR, thus led the 510-bp fragment to have the most redundancy. However, it is difficult to explain why no clone was from chromosome 1. In fact, only 8 bases were different between the sequences from mtDNA and chromosome 1, and there was a mutation of T to C, which introduced an extra CpG site in chromosome 1. Cross *et al* illuminated that about 10% of CGI clones were rDNAs (repeat units of the ribosomal RNA genes) and less than 10% were from mitochondria (*25*). Analysis of other most redundant CGI clones supported this observation; for example, HsCGICLT014311 and HsCGICLT017988 were from rDNA, while HsCGICLT005628 and HsCGICLT016746 may be from mitochondria. The reason why all of the 349 most redundant clones were derived from mtDNA and the redundancy was much higher than others (Table 1) is yet to be further investigated.

Phylogenetic analysis of six NUMTs with mitochondrial sequences from a wide range of species enabled us to identify three insertion events that occurred at different times along evolution process, which suggested that the transfer of mtDNA into nuclear genome is a continuing process. Earlier insertion events allowed the NUMTs to evolve through duplication events, making them less similar to mtDNA than those newly acquired NUMTs that were derived from recent insertion events. Therefore, NUMTs, as molecular fossils of mtDNA in the nuclear genome, could be informative for us to investigate the evolution of nuclear genome during long evolution process.

Although we do not know why this fragment of mtDNA has the highest redundancy, we do believe that the prevalent existence of such specific sequence in human CGI library suggests it may play a significant role in DNA methylation. Xie *et al* (*23*) has demonstrated that the absence of mtDNA could induce DNMT1 expression, which was responsible for the changes of methylation patterns, so we hypothesized that the regulation may depend on the combination of such specific methylated regions in mtDNA with MBDs. It is proposed that MBDs can partially mediate the linking of DNA methylation and histone modification (*29*). It is also observed that mtDNA was methylated in cancers (*30*). Therefore, it is possible that the methylated mtDNAs can be bound by the primary sequence of MBD, which may disturb the process of MBD entering the nucleus, and decrease or loss of MBDs in nucleus may alter the patterns of DNA methylation. However, further investigation will be needed to test the hypothesis and provide more clues to elucidate the mechanism of mtDNA regulating nuclear DNA methylation.

Conclusion

In our study, we observed that the most redundant sequences of 349 clones in a human CGI library were derived from mtDNA. Further analysis indicated that there was a 5,845-bp DNA transfer from mtDNA to chromosome 1, and all the clones should be the products of an MseI fragment with the length of 510 bp, which contained a putative CGI of 270 bp. Phylogenetic analysis allows us to detect two primary DNA transfer events as well as possible secondary transfer events between different chromosomes. These results may further our understanding of how the mtDNA regulates DNA methylation in the nucleus.

Materials and Methods

CGI library clones

All the clones were derived from the CGI library constructed by Wellcome Trust Sanger Institute (Sanger), which had been prepared and described by Cross *et al* (*25*). The first set of 17,606 clones (BIG18K) was sequenced by Beijing Institute of Genomics, CAS (http://methycancer.genomics.org.cn). The second set of 12,192 clones (Sanger12K) was obtained from Sanger, which was sequenced previously at Sanger and was publicly available (http://www.sanger.ac.uk/HGP/cgi.shtml). The third set of 8,373 clones (UHN8K) was downloaded from UHN

Human CpG Island Microarray Database (http://data.microarrays.ca/cpg/).

Sequence alignment and clustering of genomic loci

BLAT (Blast-like Alignment Tool) was obtained from UCSC Genome Bioinformatics (http://genome.ucsc.edu/). BLAST and the March 2006 Build (Hg18, Build 36.2) of the human genome were obtained from the NCBI (http://www.ncbi.nlm.nih.gov/). The genome is formatted for local BLAT and BLAST alignments. We used BLAT (with "-fine" option) on all the clones to the human genome first. If a clone without BLAT hit on human chromosome, or the length of insertion/deletion is more than 50 bp, or over 10% of clone length, we will do BLAST (with "-m8 -e 1e-6" option). Alignment of two sequences was carried out by bl2seq, while alignment of multiple sequences was performed using ClustalW (*31*) with the default settings. Clones sharing the same genomic locus are clustered. For each locus, one or more representative non-redundant clones are selected based on the aligned length on the chromosome.

CGI prediction and MseI fragment identification

We predicted CGIs on the $NUMT^{ND\text{-}COX}$ using CpGi130 (*27*) with criteria as follows: (1) the length of CGI no less than 200 bp, (2) GC content at least 50%, and (3) the ratio of observed CpG frequency over the expected frequency exceeds 0.60. The identification of the MseI fragments was carried out by in-house program coded by Perl language.

Phylogenetic analysis

The phylogenetic analysis was carried out using six different human NUMTs having COXII homologous sequences. One human NUMT was Chr1 6kb, which was the homologous sequence of $NUMT^{ND\text{-}COX}$ in Chr1. The other five NUMTs from Chr2, 4, 7, 11 and 17 are derived from Blastn search against "Human genomic + transcript" using Chr1 6kb, and GenBank accession number in each parenthesis indicates the according genomic contig or scaffold (Figure 4). All the sequences were multialigned and phylogenetic tree was constructed with neighbor-joining method (Kimura 2-parameter nucleotide model) using MEGA 3.1 (*32*). Only the conserved sequences of the NUMTs and their corresponding sequences in the COXII genes were used to construct the phylogenetic tree. Bootstrap percentage values are also shown in Figure 4 (n = 500 replicates).

Authors' contributions

XH and ST carried out the study and prepared the manuscript. JJ helped with data collection and analysis of alignment. SH and JY supervised the research and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1 Clark, S.J. and Melki, J. 2002. DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene* 21: 5380-5387.

2 Futscher, B.W., *et al.* 2002. Role for DNA methylation in the control of cell type specific maspin expression. *Nat. Genet.* 31: 175-179.

3 Bird, A.P. and Wolffe, A.P. 1999. Methylation-induced repression—belts, braces, and chromatin. *Cell* 99: 451-454.

4 Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev.* 16: 6-21.

5 Jaenisch, R. and Bird, A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 33 Suppl: 245-254.

6 Li, E., *et al.* 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69: 915-926.

7 Ferguson-Smith, A.C. and Surani, M.A. 2001. Imprinting and the epigenetic asymmetry between parental genomes. *Science* 293: 1086-1089.

8 Reik, W. and Walter, J. 2001. Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.* 2: 21-32.

9 Walsh, C.P. and Bestor, T.H. 1999. Cytosine methylation and mammalian development. *Genes Dev.* 13: 26-34.

10 Bird, A.P. 1978. Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of

methylated sites supports semi-conservative copying of the methylation pattern. *J. Mol. Biol.* 118: 49-60.
11 Gruenbaum, Y., *et al.* 1981. Methylation of CpG sequences in eukaryotic DNA. *FEBS Lett.* 124: 67-71.
12 Bird, A.P. 1986. CpG-rich islands and the function of DNA methylation. *Nature* 321: 209-213.
13 Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* 196: 261-282.
14 Takai, D. and Jones, P.A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* 99: 3740-3745.
15 Jones, P.A. and Baylin, S.B. 2002. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* 3: 415-428.
16 Cross, S.H., *et al.* 1994. Purification of CpG islands using a methylated DNA binding column. *Nat. Genet.* 6: 236-244.
17 Heisler, L.E., *et al.* 2005. CpG Island microarray probe sequences derived from a physical library are representative of CpG Islands annotated on the human genome. *Nucleic Acids Res.* 33: 2952-2961.
18 Jones, P.A. and Baylin, S.B. 2007. The epigenomics of cancer. *Cell* 128: 683-692.
19 Esteller, M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* 8: 286-298.
20 He, X., *et al.* 2008. MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.* 36: D836-841.
21 Lopez, J.V., *et al.* 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J. Mol. Evol.* 39: 174-190.
22 Richly, E. and Leister, D. 2004. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21: 1081-1084.
23 Xie, C.H., *et al.* 2007. Mitochondrial regulation of cancer associated nuclear DNA methylation. *Biochem. Biophys. Res. Commun.* 364: 656-661.
24 Smiraglia, D.J., *et al.* 2008. A novel role for mitochondria in regulating epigenetic modification in the nucleus. *Cancer Biol. Ther.* 7: 1182-1190.
25 Cross, S.H., *et al.* 1999. Isolation of CpG islands from large genomic clones. *Nucleic Acids Res.* 27: 2099-2107.
26 Altschul, S.F., *et al.* 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
27 Takai, D. and Jones, P.A. 2003. The CpG island searcher: a new WWW resource. *In Silico Biol*. 3: 235-240.
28 Tourmen, Y., *et al.* 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 80: 71-77.
29 Cedar, H. and Bergman, Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* 10: 295-304.
30 Maekawa, M., *et al.* 2004. Methylation of mitochondrial DNA is not a useful marker for cancer detection. *Clin. Chem.* 50: 1480-1481.
31 Thompson, J.D., *et al.* 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* Chapter 2: Unit 2.3.
32 Kumar, S., *et al.* 2004. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief. Bioinform.* 5: 150-163.

Supplementary Material

Table S1 and Figure S1

DOI: 10.1016/S1672-0229(10)60009-5